Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 70 (2015) 265 – 273

Procedia Computer Science

4th International Conference on Eco-friendly Computing and Communication Systems

Development of a Java Library for Protein Stability and Disorder Computations

Subrata Sinha[a*], Bishwajit Bora[a], G.C Hazarika[b]

[a]*Centre for Bioinformatics Studies, Dibrugarh University,Dibrugarh-786004,India*
[b]*Department of Mathematics, Dibrugarh University,Dibrugarh-786004,India*

Abstract

Faster, reliable and accurate development is an important issue in any field of software development supported by extended language libraries. Structural bioinformatics software development faces tremendous challenges while developing software tools and utilities because such type of development needs strong knowledge on structural biology and programmers take tremendous pressure to write the code from scratch increasing the development time, cost of the software leading to overall TOC of software. Existing libraries Bio Java and MESHI lacks those functionalities which are most frequently expected by the developers of this domain like inbuilt functions to predict the total energy of a protein structure contributed by various bonds, to predict the structural stability of a protein based on torsion angles, finding disordered clusters and disorder regions from protein structure are few of them. In this paper an attempt has been made to develop a library pBio which can predict the structural stability of protein in terms of torsion angles. It can Calculate Phi/Psi angle for all residues, identifying disordered residues, disorder clusters, percentage of disordered residues, residues involving in various types of bonds, calculations of bond energy and such functionalities can be further used in the problems like active site prediction based on disorder regions, protein energy minimization.

Peer-review under responsibility of the Organizing Committee of ICECCS 2015

Keywords: Structural Stability, Disorder region, disorder clusters, Active Site Prediction Structural Bioinformatics.

* Corresponding author. Tel.: +91-848-646-8313
E-mail address: subratasinha@dibru.ac.in

Peer-review under responsibility of the Organizing Committee of ICECCS 2015
doi:10.1016/j.procs.2015.10.086

1. Introduction

Protein stability knowledge is essential to the understanding of their structure and function. The overall structure of proteins can be described by their backbone dihedral angles. Main-chain configuration entropy is a strong force opposing the stability of native proteins (Dill , 1990) and can be visualized by generating Ramachandran plots, which show the correlations between the ϕ(Phi) , ψ(Psi) dihedral angles (Ramachandran *et al.*, 1963) . Ramachandran showed that the percentage of ϕ and ψ angles of the amino acid residues in favoured region of Ramachandran plot is a determinant of protein stability. Bio Java lacks the function to compute the percentage of ϕ and ψ angles of the amino acid residues in favoured region of Ramachandran plot. In addition, clustering in favored region is very significant when comparing stability of two protein molecules, which have same percentage of ϕ and ψ angles of the amino acid residues in favoured region of Ramachandran plot. Tight clustering indicates more stability and loose clustering indicates less stability. Bio Java (Prlic, A *et al*, 2012) has not incorporated this important feature by which we can predict the structural stability in terms of clusteredness in favoured region of Ramachandran plot. Moreover, the getPhi() and getPsi() functions of Calc class in BioJava are not compatible when the dataset to be analyzed is large , since , these functions are unable to calculate ϕ and ψ angles of all the residues of a protein at once , the programmer needs to provide two consecutive residues iteratively to these functions , which takes much time and thus these functions are not efficient in dealing with large dataset .

Disulphide bonds are well known to play key roles in stability, folding and functions of proteins (Ratna R Thangudu *et al.*, 2008, Omid Ranaei Siadat *et al.*, 2006). Disulfide bonds stabilize the native conformation of a protein by decreasing conformational entropy (Philip J. Hogg, 2003). Arnold Mcauley *et al.* (2007) suggested that a disulphide bond has 50 - 75 kcal/mol of dissociation energy. According to Sinan Ketena, *et al.* (2012) and Pace *et al.* (1988), the increase in the stability of the native structure due to the formation of a particular disulfide bond is directly proportional to the number of residues between the linked cysteines, the larger the number of residues between the disulfide, the greater is the stability imparted to the native structure. Bio Java has a function getSSBond() to calculate the number of disulfide bonds formed in a protein , but this function is unable to measure the disulfide bond dissociation energy which is very crucial in protein stability . Thus , there is a strong need of such a function that can calculate the total number of disulfide bonds in a protein as well as the dissociation energy of those bonds .

Some proteins or particular regions of proteins lack a well-defined tertiary structure in their native state. Such proteins are called as intrinsically disordered proteins and likewise the unstructured regions are called as intrinsically disordered protein regions (A.K. Dunker *et al.*, 2001). Disordered residues or regions have great significance in protein stability. Disorder provides the basis for numerous functions, but an especially interesting one is the involvement of disorder in molecular recognition which include include enzyme-substrate, receptor-ligand, protein-peptide, protein-protein, protein-RNA and protein-DNA complexes.

The identification of ligand-binding sites is often the starting point for protein function annotation and structure- based drug design. In a study of lysozyme crystals, Artymiuk P J *et al.* (1979) has revealed that the residues of highest apparent motion, i.e., disordered residues are in the active site region, where conformational change has been observed upon substrate binding. According to Altman *et al.* 1994, disordered residues form clusters when come into close contact with each other and these clusters may participate in ligand-binding. BioJava provides a class Jronn, for analysis of protein disorder from a protein sequence. But this class has some limitations, such as - the input sequence must not contain any ambiguous character, and have a minimum length of 19 amino acids. Moreover , Bio Java avoids the structural information of proteins in analyzing disorder , it only considers the sequence information which is not adequate and must be resolved.

Moreover , extensive studies of literature has suggested that these two , i.e., protein stability and disorder are correlated . Fitzkee NC et *al.* (2008) concluded that disordered residues prevent the formation of stable tertiary structures in proteins.

So we developed an open source library for protein stability and disorder which includes the methods – getAllPhiPsi(), getPercentage(), getDispersion() in the class Torsion, getDisorderedResidues(), getDisorderedRegions(), getDisorderedClusters() in the class Disorder and getDisulfideBonds() in the class Bond that are not available in Bio Java but frequently needed by the programmers in the development of structural bioinformatics software.

2. Materials and Method

2.1. Material

JDK 1.8.0 , Protein Data Bank , 458 PDB structures (X-ray crystallography , resolution <= 1.2 Å) of eukaryotic proteins to test our library, RAMPAGE Ramachandran Plot Analysis and DIHED2 to compare results of Torsion class methods, Bio Java Library.

2.2 Method

Torsion Class: (a) getAllPhiPsi() method accepts a PDB file as input parameter and calculates phi/psi angles from the atomic coordinate information. (b) getPercentage() method calculates the percentage of ϕ and ψ angles of the amino acid residues in favoured region of Ramachandran plot and (c) getDispersion() calculates the scatteredness of ϕ and ψ angles plots in the favoured region of Ramachandran plot.

Disorder class: (a) getDisorderedResidues() returns all residues whose B-factor is > 50 and whose coordinates are missing in the PDB, (b) getDisorderedRegions() returns segments of consecutive disorder residues along with their starting and ending position in the sequence, (c) getDisorderedClusters() returns the disordered residues within a range of 5 *Å* in 3D space.

Bonds class: (a) getDisulfideBonds() method returns the all bonds formed between CYS-CYS residues along with their bond energy in Kcal/mol.

3. Results

3.1 Result of getAllPhiPsi() method of Torsion class : getAllPhiPsi() method has been executed for 458 PDB Structures and phi/psi pair values has been validated with RAMPAGE as well as DIHED2, the result shows that the values returned by the method are matching with these two standard tools. The comparative charts of ϕ and ψ values for the protein Interleukin 13 (PDB ID: 3BPO) are given as below in Table 3.1.1 and 3.1.2 respectively.

Table 3.1.1 Comparative Chart of Phi/Psi values of RAMPAGE and getAllPhiPsi() method for the protein *Interleukin 13*

RAMPAGE RESULT			**getAllPhiPsi() RESULT**		
Identifier	**Φ**	**Ψ**	**Identifier**	**ϕ**	**ψ**
A:9:ALA	-59.24	-68.60	A:9:ALA	-59.24	-68.60
A:27:PRO	-101.25	138.05	A:27:PRO	-101.25	138.05
A:29:CYS	79.37	38.58	A:29:CYS	79.37	38.58
A:37:ILE	-127.53	58.20	A:37:ILE	-127.53	58.20

A:42:GLY	126.36	-67.51	A:42:GLY	126.36	-67.51
A:52:ILE	-59.81	4.93	A:52:ILE	-59.81	4.93
A:55:SER	-108.45	-69.07	A:55:SER	-108.45	-69.07
A:56:GLY	-129.93	53.02	A:56:GLY	-129.93	53.02
A:57:CYS	-110.84	82.26	A:57:CYS	-110.84	82.26
A:83:LEU	-79.14	-161.12	A:83:LEU	-79.14	-161.12
A:85:VAL	-149.46	75.34	A:85:VAL	-149.46	75.34
A:87:ASP	-120.84	-75.05	A:87:ASP	-120.84	-75.05
A:89:LYS	-153.12	91.33	A:89:LYS	-153.12	91.33
B:0:PRO	-65.77	80.92	B:0:PRO	-65.77	80.92
B:14:MET	-138.45	-76.12	B:14:MET	-138.45	-76.12
B:15:SER	-135.84	-45.05	B:15:SER	-135.84	-45.05
B:24:ASN	-93.13	36.76	B:24:ASN	-93.13	36.76
B:41:PHE	-68.84	75.17	B:41:PHE	-68.84	75.17
B:42:LEU	-63.78	81.76	B:42:LEU	-63.78	81.76
B:54:ASN	-129.61	57.73	B:54:ASN	-129.61	57.73

Table 3.1.2 Comparative Chart of Phi/Psi values of DIHED2 and getAllPhiPsi() method for the protein Interleukin 13

DIHED2 RESULT					**getAllPhiPsi() RESULT**				
Chain Id	**Res. No.**	**Res. Name**	**ϕ**	**ψ**	**Chain Id**	**Res. No.**	**Res. Name**	**Φ**	**ψ**
A	2	GLY	999.99	-160.48	A	2	GLY	0	-160.48
A	3	PRO	-87.93	14.16	A	3	PRO	-87.93	14.16
A	4	VAL	-103.88	149.3	A	4	VAL	-103.88	149.3
A	5	PRO	-61.03	139.45	A	5	PRO	-61.03	139.45
A	6	PRO	-62.6	-43.16	A	6	PRO	-62.6	-43.16
A	7	SER	-46.38	-40.12	A	7	SER	-46.38	-40.12
A	8	THR	-61.47	-33.38	A	8	THR	-61.47	-33.38
A	9	ALA	-59.24	-68.6	A	9	ALA	-59.24	-68.6
A	10	LEU	-55.89	-36.72	A	10	LEU	-55.89	-36.72
B	1	PHE	-57.48	124.61	B	1	PHE	-57.48	124.61
B	2	LYS	-147.56	159.06	B	2	LYS	-147.56	159.06
B	3	VAL	-103.9	103.28	B	3	VAL	-103.9	103.28
B	4	LEU	-78.93	-26.89	B	4	LEU	-78.93	-26.89
B	5	GLN	-130.66	112.58	B	5	GLN	-130.66	112.58
B	6	GLU	-51.38	129.32	B	6	GLU	-51.38	129.32
B	7	PRO	-51.33	122.03	B	7	PRO	-51.33	122.03
B	8	THR	-101.12	164.58	B	8	THR	-101.12	164.58
B	9	CYS	-155.18	147.26	B	9	CYS	-155.18	147.26
B	10	VAL	-136.92	157.48	B	10	VAL	-136.92	157.48
C	32	GLN	999.99	92.06	C	32	GLN	0	92.06
C	33	PRO	-73.61	168.72	C	33	PRO	-73.61	168.72
C	34	PRO	-74.9	165.62	C	34	PRO	-74.9	165.62
C	35	VAL	-78.19	140.71	C	35	VAL	-78.19	140.71
C	36	THR	-76.29	138.3	C	36	THR	-76.29	138.3
C	37	ASN	42.76	58.24	C	37	ASN	42.76	58.24
C	38	LEU	-56.45	125.97	C	38	LEU	-56.45	125.97
999.99 = Could not calculate the angle					0 = Could not calculate the angle				

3.2 Result of getPercentage() method of Torsion class: getPercentage() method has been executed for 458 proteins and the results of 10 proteins - Tumor Necrosis Factor-α (PDB ID : 1A8M) , Interleukin-4 (PDB ID : 1BBN) , Eotaxin (PDB ID : 1EOT) , Interleukin-3 (PDB ID : 1JLI) , Interleukin-13 (PDB ID : 3BPO) , Interleukin-5 (PDB ID : 3VA2) , Hemoglobin (PDB ID : 1A00) , Cu/Zn Superoxide Dismutase (PDB ID : 1B4T) , Trypsin (PDB ID : 1C1N) , and Nitric Oxide Synthase (PDB ID : 1DM6) have been given in table 3 . The table 3.2.1 contains percentage of phi and psi angles in favoured, allowed and disallowed regions for all residues as a whole.

Table 3.2.1 Results of getPercentage() method for proteins of PDB ID - 1A8M , 1BBN , 1EOT , 1JLI , 3BPO , 3VA2 , 1A00 , 1B4T , 1C1N , and 1DM6 .

PDB ID	% In Favoured Region	% In Allowed Region	% In Disallowed Region
1A8M	77.41	13.82	8.77
1BBN	81.95	7.52	10.53
1EOT	81.08	8.11	10.81
1JLI	76.79	11.61	11.61
3BPO	79.04	10.65	10.31
3VA2	79.54	9.37	11.09
1A00	89.72	4.01	6.27
1B4T	71.9	9.8	18.3
1C1N	78.03	10.76	11.21
1DM6	82.65	7.59	9.76

3.3 Result of getDispersion() method of Torsion class : getDispersion() method has been executed for 458 proteins and the dispersion in favoured region for 3VA2 and 1A00 is found to be 58.89 and 61.28 which is supported by the Ramachandran plots generated by RAMPAGE as in Fig. 3.3(a) and 3.3(b) respectively.

3(a)

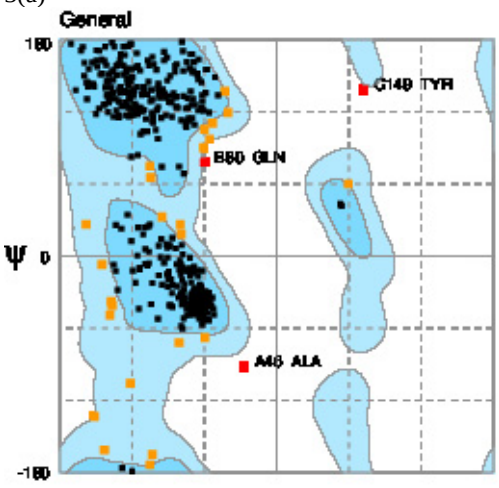

Figure 3.3(a). Ramachandran Plot of PDB ID 3VA2

3(b)

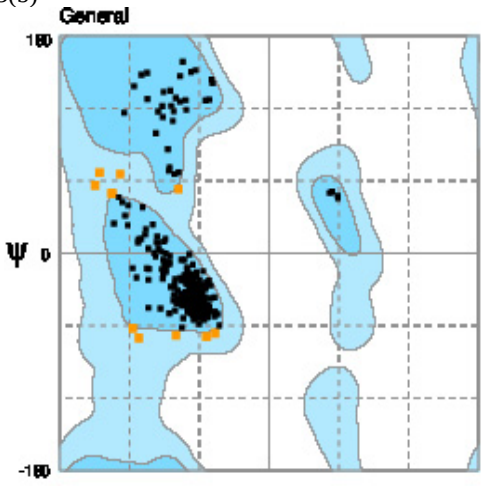

Figure 3.3(b). Ramachandran Plot of PDB ID 1A00

3.4 Results Of getDisulfideBonds() Method Of Bond Class : getDisulfideBonds() method has been executed for 458 proteins and the results of Trypsin (PDB ID : 1C1N) have been given in table 3

Table 3.4.1 Chart of Cysteine residues forming disulfide bonds along with bond energy of the protein ***Trypsin*** (PDB ID : 1C1N)

Bond Forming Residues	Bond Energy (KCal/mol)	Bond Forming Residues	Bond Energy (KCal/mol)
A:22 A:157	65.99	A:136 A:201	62.50
A:42 A:58	73.50	A:168 A:182	59.87
A:128 A:232	55.63	A:191 A:220	58.26

3.4(a)

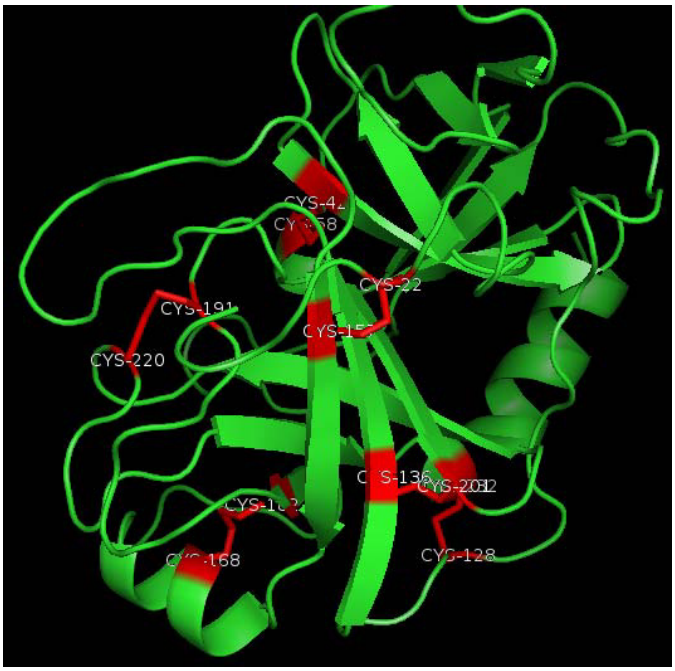

Fig. 3.4(a) Cystein residues forming disulfide bonds of the protein ***Trypsin*** (PDB ID : 1C1N) in PyMOL

3.5 Result of getDisorderedResidues() method of Disorder class : getDisorderedResidues() method has been executed for 458 proteins and the result of 3BPO is given below in table 3.5.1

.

Table 3.5.1 All disordered residues of the PDB ID : 3BPO returned by *getDisorderedResidues()* method

Chain ID	Residue Number	Residue Name	Chain ID	Residue Number	Residue Name	Chain ID	Residue Number	Residue Name
A	1	PRO	A	124	ASP	C	108	GLU
A	23	ASN	A	125	ARG	C	109	LYS
A	24	GLN	A	126	THR	C	110	PRO
A	25	LYS	B	1	PHE	C	111	SER
A	38	LEU	B	2	LYS	C	124	ASP
A	74	LYS	B	3	VAL	C	151	ASN
A	75	VAL	B	4	LEU	C	152	THR
A	76	SER	B	5	GLN	C	153	SER
A	77	ALA	B	197	HIS	C	192	VAL
A	78	GLY	B	198	ASN	C	193	LYS
A	79	GLN	B	199	SER	C	194	ASP
A	80	PHE	B	200	TYR	C	195	SER
A	81	SER	B	201	ARG	C	196	SER
A	113	ASN	B	202	GLU	C	197	PHE
A	114	ARG	C	29	THR	C	198	GLU
A	115	ASN	C	30	GLU	C	199	GLN
A	116	PHE	C	31	THR	C	267	SER

A	117	GLU	C	72	ASP	C	268	GLN
A	118	SER	C	73	LYS			
A	119	ILE	C	103	SER			
A	120	ILE	C	104	THR			
A	121	ILE	C	105	ASN			
A	122	CYS	C	106	GLU			
A	123	ARG	C	107	SER			

3.6 Results of getDisorderedRegions () method of Disorder class: getDisorderedRegions() method has been executed for 458 PDB structures and the results of 3BPO disordered regions are given below in Table 3.6.1

Table 3.6.1 All disordered regions of the PDB ID: 3BPO

Chain ID	Starting Position	Ending Position	Disordered Regions
A	74	81	LYSVALSERALAGLYGLNPHESER
A	113	126	ASNARGASNPHEGLUSERILEILEILECYSARGASPARGTHR
B	1	5	PHELYSVALLEUGLN
B	197	202	HISASNSERTYRARGGLU
C	103	111	SERTHRASNGLUSERGLULYSPROSER
C	192	199	VALLYSASPSERSERPHEGLUGLN

3.6 (a).

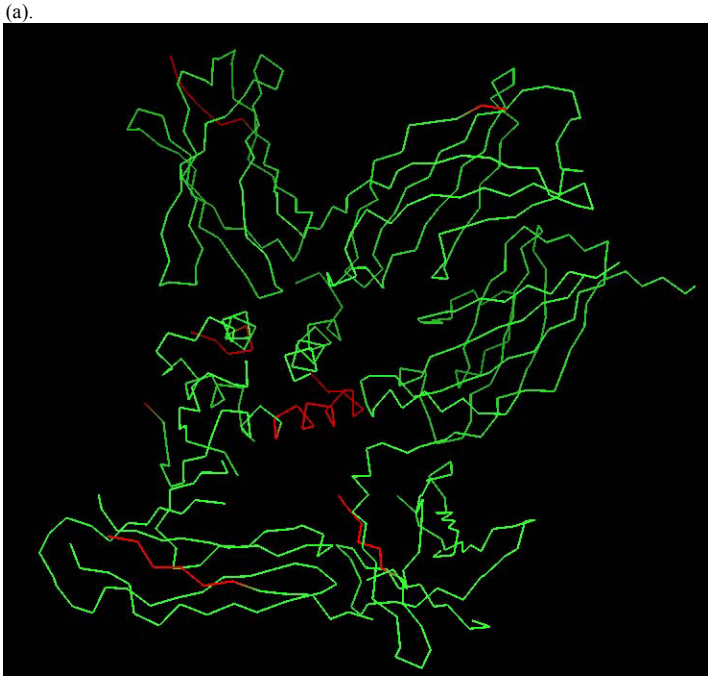

Figure 3.6 (a). Disordered regions (in red) in the protein Interleukin-13 (PDB ID: 3BPO) in PyMOL.

3.7 Results of getDisorderedClusters() method of Disorder class : getDisorderedClusters() method has been executed for 458 PDB Structures and the results of the protein Interleukin-5 (PDB ID : 3VA2) , have been given in table 3.7.1

Table 3.7.1. Disordered clusters of the protein Interleukin-5 (PDB ID: 3VA2)

Disordered Clusters	Disordered Clusters	Disordered Clusters	Disordered Clusters	Disordered Clusters
B : 112 : VAL	C : 323 : SER	B : 82 : GLY	A : 28 : ALA	C : 166: ASP
B : 113 : ASN	A : 123 : LEU	B : 69 : GLN	C : 270 : PHE	B : 84 : VAL
C : 71 : ALA	C : 132 : LEU	B : 70 : GLY	C : 271 : ASP	B : 85 : GLU
C : 72 : PRO	C : 133 : THR	C : 103 : GLN	A : 50 : LEU	B : 72 : GLY
A : 53 : PRO	A : 49 : THR	C : 104 : ASN	B : 89 : LYS	B : 73 : THR
B : 109 : ARG	C : 329 : ILE	C : 320 : GLY	B : 90 : ASN	C : 217 : GLY
C : 272 : TYR	C : 330 : TYR	C : 321 : LEU	C : 42 : GLY	C : 218 : SER
C : 273 : GLU	A : 95 : LYS	C : 174 : ARG	C : 43 : LEU	B : 50 : LEU
C : 99 : ARG	A : 96 : LYS	C : 175 : TYR	C : 171 : LEU	B : 51 : ARG
C : 100 : THR	C : 115 : SER	B : 31 : LYS	C : 172 : TYR	A : 92 : SER
C : 47 : LEU	C : 116 : ALA	B : 32 : GLU	B : 47 : ASN	A : 93 : LEU
C : 48 : LEU	B : 110 : ARG	A : 69 : GLN	B : 48 : GLU	A : 107 : GLU
C : 113 : TRP	B : 111 : ARG	C : 245 : VAL	B : 126: MET	A : 108 : GLU
C : 114 : ALA	C : 298 : ILE	C : 246 : THR	B : 127 : ASN	C : 130 : VAL
C : 185 : TYR	C : 299 : ASP	B : 48 : GLU	C : 101 : ILE	
C : 186 : SER	C : 67 : VAL	A : 122 : PHE	C : 102 : LEU	
C : 287 : GLU	C : 176 : GLY	C : 242 :PRO	C : 145 : ARG	
C : 288 : LYS	C : 177 : SER	C : 243 : LEU	C : 146 : LEU	
C : 322 : TRP	B : 81 : GLY	A : 27 : SER	C : 165 : GLU	

3.7 (a).

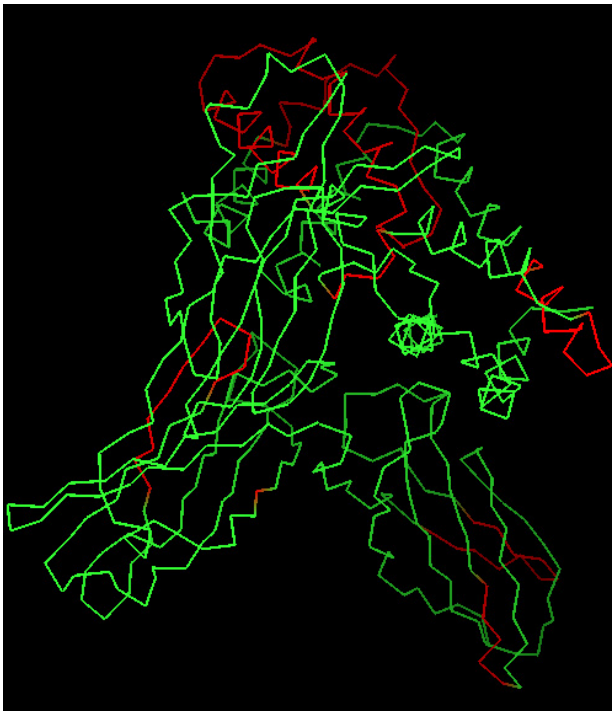

Figure 3.7 (a). Visualization of disordered clusters in the protein Interleukin-5 (PDB ID : 3VA2) in PyMOL

4. Conclusion

Development of methods which are unavailable in present libraries for structural bioinformatics will help the developers to develop fast, reliable utilities. The proposed library can be further extended to calculate total energy of a protein structure contributed by various bonds like hydrogen bond, Salt Bridge, weak interaction forces in the PDB structure. The bond class can be very useful for calculating total energy

contributed by the various bonds in a protein structure; hence such method can be very much useful while developing protein energy minimization software. The **getDispersion()** method of Torsion class can be very much useful while bulk assessment of protein stability based on Ramachandran plot. Methods of Disorder class of the developed library can be further extended and made useful for active site prediction software.

Acknowledgements

We acknowledge Prof. Subrata Chakraborty, Director i/c Centre for Bioinformatics Studies for facilitating with the infrastructure needed to perform the experiments and we acknowledge Dr. K Narain, Dy. Director of RMRC-ICMR, Dibrugarh for his constant support and help.

References

1. Ken A. Dill(1990); Dominant Forces in Protein Folding; Biochemistry, Vol. 29(31):7133-7151
2. Ramachandran G. N. , Ramakrishnan C. , Sasisekharan V.(1963) ; Stereochemistry of Polypeptide Chain Configurations; *J. Mol. Biol.* 7,95-99
3. A. Keith Dunker and Zoran Obradovic(2001); The protein trinity—linking function and disorder ; *Nature Biotechnology* Vol.19
4. Artymiuk PJ, Blake CC, Grace DE, Oatley SJ, Phillips DC, Sternberg M. (1979); Crystallographic studies of the dynamic properties of lysozyme*; Nature.* 280(5723):563-8.
5. Altman, R.B., Hughes, C., and Jardetzky, O. (1994). Compositional characteristics of disordered regions in proteins*. Prot. Pept. Lett.* 2:120–127.
6. Nicholas C. Fitzkee And Bertrand Garci´A-Moreno E.; Electrostatic effects in unfolded staphylococcal nuclease; *Protein Science* (2008), 17:216–227
7. Ratna R Thangudu,Malini Manoharan, N Srinivasan, Frédéric Cadet, R Sowdhamini and Bernard Offmann; Analysis on conservation of disulphide bonds and their structural features in homologous protein domain families; *BMC Structural Biology* 2008, 8:55
8. Omid Ranaei Siadat Andrée Lougarre, Lucille Lamouroux,Caroline Ladurantie and Didier Fournier (2006); The effect of engineered disulfide bonds on the stability of Drosophila melanogaster acetylcholinesterase; *BMC Biochemistry* , 7:12
9. Philip J. Hogg(2003); Disulfide bonds as switches for protein function; *TRENDS in Biochemical Sciences* Vol.28 (4): 210-214
10. Arnold Mcauley, Jaby Jacob, Carl G. Kolvenbach, Kimberly Westland, Hyo Jin Lee, Stephen R. Brych, Douglas Rehder, Gerd R. Kleemann, David N. Brems, And Masazumi Matsumura; Contributions of a disulfide bond to the structure,stability, and dimerization of human IgG1 antibody CH3 domain; *Protein Science* (2008), 17:95–106
11. Sinan Ketena,Chia-Ching Choua, Adri C.T. van Duinc, Markus J. Buehlera; Tunable nanomechanics of protein disulfide bonds in redox microenvironments; *Journal Of The Mechanical Behavior Of Bio Medical Materials* 5 (2012): 32 – 40
12. C. Nick Pace, Gerald R. Grimsley, James A. Thomson, and Ben J. Barnett; Conformational Stability and Activity of Ribonuclease TI with Zero,One, and Two Intact Disulfide Bonds(1988); *The Journal Of Biological Chemistry*; Vol. 263(24):11620-11825
13. Prlic, A. Yates, S.E. Bliven, P.W. Rose, J. Jacobsen, P.V. Troshin, M. Chapman, J. Gao, C.H. Koh, S. Foisy, R. Holland, G. Rimsa, M.L. Heuer, H. Brandstatter-Muller, P.E. Bourne, S. Willis, BioJava: an open-source framework for bioinformatics in 2012, *Bioinformatics* 28 (20) (Oct 15 2012) 2693–2695